Rethinking with Retrieval: Faithful Large Language Model Inference

Year: 2022 | *Citations: 198* | *Authors: Hangfeng He, Hongming Zhang, D. Roth*

Abstract

Despite the success of large language models (LLMs) in various natural language processing (NLP) tasks, the stored knowledge in these models may inevitably be incomplete, out-of-date, or incorrect. This motivates the need to utilize external knowledge to assist LLMs. Unfortunately, current methods for incorporating external knowledge often require additional training or fine-tuning, which can be costly and may not be feasible for LLMs. To address this issue, we propose a novel post-processing approach, rethinking with retrieval (RR), which retrieves relevant external knowledge based on the decomposed reasoning steps obtained from the chain-of-thought (CoT) prompting. This lightweight approach does not require additional training or fine-tuning and is not limited by the input length of LLMs. We evaluate the effectiveness of RR through extensive experiments with GPT-3 on three complex reasoning tasks: commonsense reasoning, temporal reasoning, and tabular reasoning. Our results show that RR can produce more faithful explanations and improve the performance of LLMs.